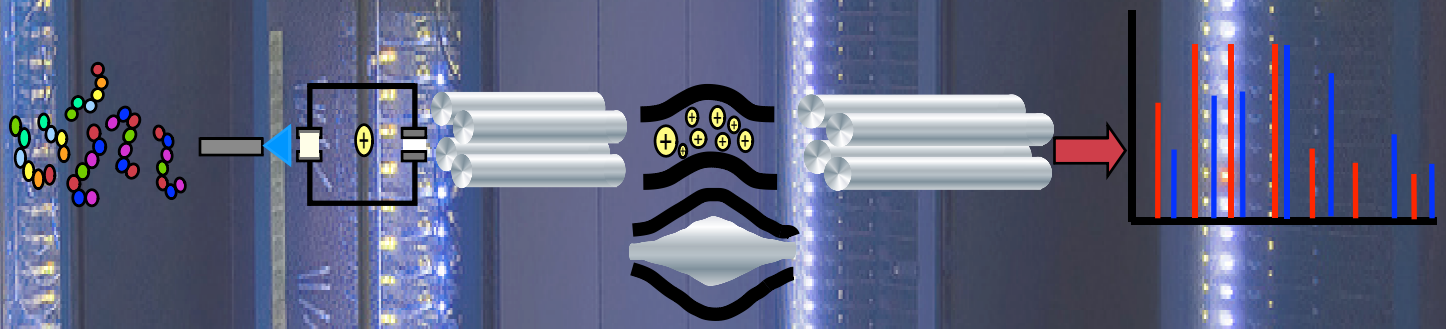
Mass Spectrometry Informatics Big Data to Knowledge

A Report from an NSF Supported Workshop Held May 11-12, 2015 in Arlington, VA



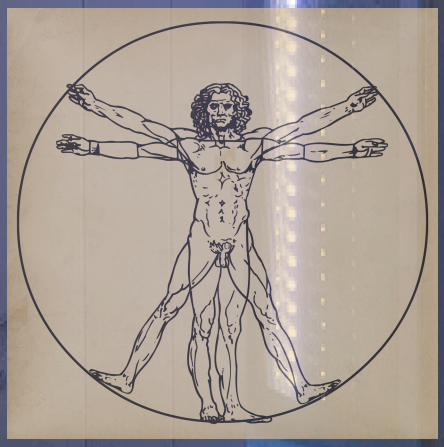


TABLE OF CONTENTS

EXECUTIVE SUMMARY

Proteins occupy a central position in molecular biology, connecting genomics and transcriptomics to the cellular structures, functions, signals, and metabolism that underpin simple and complex phenotypic traits in any species in any circumstance. The systematic study of proteins is crucial for understanding biological systems. For two decades the field of proteomics (large-scale protein analysis) has been driven by advances in mass spectrometry instrumentation, supplemented by sophisticated separation methods, allowing researchers to simultaneously identify the many thousands of proteins comprising the proteome. The analysis of proteomic mass spectrometry data by efficient and accurate database searching methods, relying heavily on mathematical models and computer algorithms, represented a key breakthrough for the field to enable large-scale data acquisitions. What has become clear from 20 years of proteomics and mass spectrometry research is the synergistic relationship between instrument and software development. The net effect of these new tools, in synergy with other 'omics' technologies, has been a rapid acceleration of the pace of biological research.

Participants of the National Science Foundation (NSF) sponsored workshop *Mass Spectrometry: Data to Knowledge, Arlington, VA, May 11-12, 2015* met to discuss challenges, opportunities, and research needs for computational proteomics. The workshop's goal was to understand the computational, algorithmic and statistical methods needed to drive collection of the best biological information from large-scale mass spectrometry datasets over the next 10-15 years. The discussions identified grand challenges and methodology gaps in three major areas of proteomics: (1) instrument design and data production, (2) proteomic algorithms and computation, and (3) translating data to knowledge through integration with the existing research infrastructure.

Proteomic scientists addressed the type of experiments used to answer specific biological questions, how and what type of data is produced, and what biological knowledge is created, along with algorithms and computational methods for the effective and efficient analysis of proteomics data, including statistical analyses used to ensure sound and accurate data. The workshop identified weaknesses in methodology, the need for new approaches to get desired information, and bottlenecks to creating new biological information or insights from the data. Forward-thinking discussions identified suggestions for programs to create new mathematical and algorithmic frameworks to advance the use of mass spectrometry-based proteomics in research across plants, bacteria, archaea and newly-emerging model organisms.

GRAND CHALLENGES AND RECOMMENDATIONS

The workshop explored and discussed challenges for the future of proteomics mass spectrometry and proteomics informatics. Important challenges were identified, ranging from predictive and reactive software for data acquisition to integration of proteomic results with the biological data infrastructure. Thus significant opportunities were identified to move the field forward over the next 10 to 15 years, which should help drive future discoveries in molecular biology and help drive new types of clinical diagnostics.

1) **Software Tools and Algorithms.** Software to interpret data created by mass spectrometers can help drive the development and adoption of new experimental methods. The ability to dissociate peptide and proteins ions in a mass spectrometer provides a means to create information about the covalent structure of these molecules. Recent advances in instrument design have created new hybrid mass spectrometers with capabilities for Collision Induced Dissociation (CID), Higher-Energy Collision Dissociation (HCD), Electron Transfer Dissociation (ETD), Surface Induced Dissociation (SID), Infrared Multi-Photon Dissociation (IRMPD), and Ultraviolet Photodissociation (UVPD). These highly effective methods to dissociate ions can be used serially or in combination to create new data acquisition paradigms for more systematic and comprehensive collection of data from peptides, intact proteins and whole protein complexes. For example, multiplexed and multi-modal multi-stage mass spectrometry (MS^n) methods that collaboratively and cooperatively work with computational methods could be developed. Thus, new opportunities to develop software tools and algorithms for innovative and systematic data acquisition paradigms are possible by integrating the development of new mass spectrometry methods with software development.

2) **Parallelized data Acquisition Strategies.** A fundamental deficiency of mass spectrometry is the serial nature of data acquisition and sample analysis, so parallelized data acquisition strategies are commonly employed to increase throughput and efficiency in data acquisition. Through a combination of novel data acquisition strategies and the software tools to interpret the data, new strategies to multiplex data acquisition strategies for peptides, intact proteins and whole protein complexes could be possible. Multiplexing samples through the use of isobaric covalent tagging methods has created enormous opportunities for biological analyses. Can other forms of multiplexing through the use of sophisticated statistical analyses, predictive data modeling, and machine-learning techniques be created? Multiplexing data acquisition and sample analysis could greatly improve throughput,

efficiency and scale of mass spectrometry experiments. Thus, strategies to create new ways to massively parallelize mass spectrometry analyses could create new experimental economies for proteins biochemistry in much the same way as next generation DNA sequencing has done for genomics.

3) **New Data Formats and Compression Methods.** A general trend over the last decade has been an increase in mass spectrometer scan speeds by ~2-5 times every 2 years. As a result, mass spectrometry data set sizes have been increasing rapidly and will continue to do so in the future. An increasing amount of data and the potential for data sets to contain denser data with the development of new types of experiments will necessitate new data file formats and data compression methods to enable the movement of these data sets among collaborators and storage facilities. Additionally, the larger data sets will require the development of high throughput and large-scale data analysis tools, and computing architectures will be essential to create durable and accurate interpretations of mass spectrometry data for peptides and intact proteoforms.

4) **Mass Spectrometry-Based Structural Biology of Endogenously Formed Complexes**. Mass spectrometry is increasingly contributing to structural biology analyses. Most current studies involve *in vitro* studies of proteins, their proteoforms, and multi-protein complexes with and without their bound ligands. A future drive to study the analogous entities formed *in vivo* will help determine their native biological structures and their dynamics over the course of biological processes. These goals will necessitate the development of robust software tools for MS-based structural biology of endogenously-formed complexes, including large-scale cross-linking experiments, residue-level specificity in hydrogen/deuterium exchange, and covalent labeling methods including oxidative footprinting. Furthermore, connecting data from these methods with other types of *in vivo* analyses will increase the demand for computational interpretation and rendering of large and complicated data sets. Thus, the development of methods to create and capture mass spectrometry-based structural information *in vivo* and to combine that data with protein structural prediction algorithms to create high-resolution predictions of native cellular protein structures and protein complexes will drive our understanding of cellular biology and biological mechanisms.

5) **Development of New Statistical Tools For Mass Spectrometry Based Analyses.** As big data emerges in mass spectrometry and proteomics, these larger data sets will present unique opportunities. Many of the challenges of big data are not related as much to the size of the data sets, but to the existence of noise and errors in the data. Larger data sets will require the development of new

statistical methods for False Discovery Rate (FDR) estimation at multiple levels of organization (spectra, peptide, proteins, organisms, higher taxa) and multiple levels of error (incorrect localization, modifications, sequence, protein family). To improve statistical analyses, methods to assess and evaluate the quality of spectra to eliminate poor quality spectra or spectra of noise or non-peptides will help this process.

6) **Integration of Mass Spectrometry Data And Interpretations with Existing Knowledge-Bases.** As the overriding goal of many experiments is to discover new biological insights, the development of tools to integrate mass spectrometry data and interpretations with existing knowledge-bases to speed data to knowledge transformations. As more biological data are collected, they are stored in databases to create knowledge bases. These databases can range from repositories of data to highly sophisticated curated knowledge bases. To speed the analysis and interpretation of proteomic data, these knowledge bases should be affiliated with tools to mine this information. Better affiliation of databases can be accomplished by creating consolidation sites that collect specific types of data from all existing databases or by creating web crawlers that search on demand using specific queries to find all available data on the web.

This workshop was supported by NSF grant 1349575. The opinions presented in this report are those of the participants and not of the National Science Foundation.

INTRODUCTION AND BACKGROUND

Proteomics is the simultaneous biochemical analysis of many proteins. Over the past two decades, several key technological advances have converged to make protein biochemistry practical on a large scale. Most important was the development of tandem mass spectrometry for sequencing proteins and protein derived peptides (1). This technique, in combination with electrospray ionization (ESI) and high-performance liquid chromatography (HPLC), enables the analysis of complex protein and peptide mixtures (2). As peptide ions enter the tandem mass spectrometer they are separated based on their mass-to-charge (m/z) values and then activated to fragment, most commonly by high-energy collisions

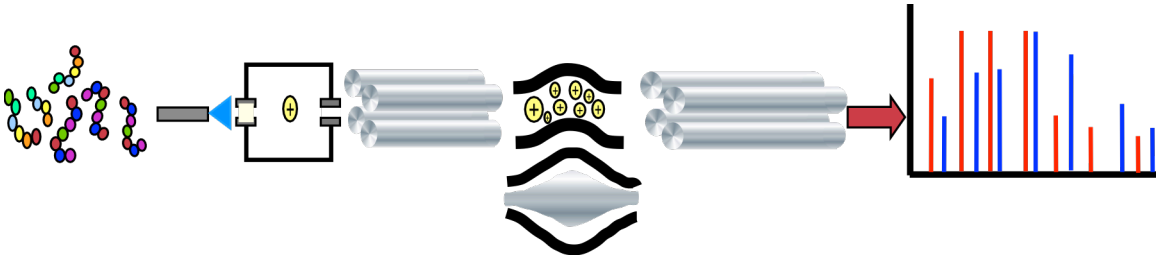


Figure 1: Peptide sequencing by tandem mass spectrometry. Peptide ions are created by electrospray ionization. M/z values for the peptides are recorded and one m/z value is selected, fragmented and the fragment ions measured. The resulting spectrum represents the amino acid sequence of the peptide.

with helium or nitrogen gas (Figure 1). The resulting fragment ions are then separated by their m/z value in the mass analyzer and counted, to form a tandem mass spectrum. The pattern of fragment ion m/z values can be used to infer the amino acid sequence of the selected peptide.

This analysis strategy requires highly efficient computer directed operation of the mass spectrometer and the development of data-dependent control of spectral acquisition was necessary to allow rapid, automated acquisition of fragmentation spectra from many peptide ions (Figure 2) (3). Instrument control software has enabled “shotgun proteomics” – a process where mixtures of proteins are enzymatically digested to create peptides mixtures, which are rapidly analyzed by liquid chromatography-tandem mass spectrometry (4-6). All modern mass spectrometers have the capability to automate data acquisition in a sophisticated manner.

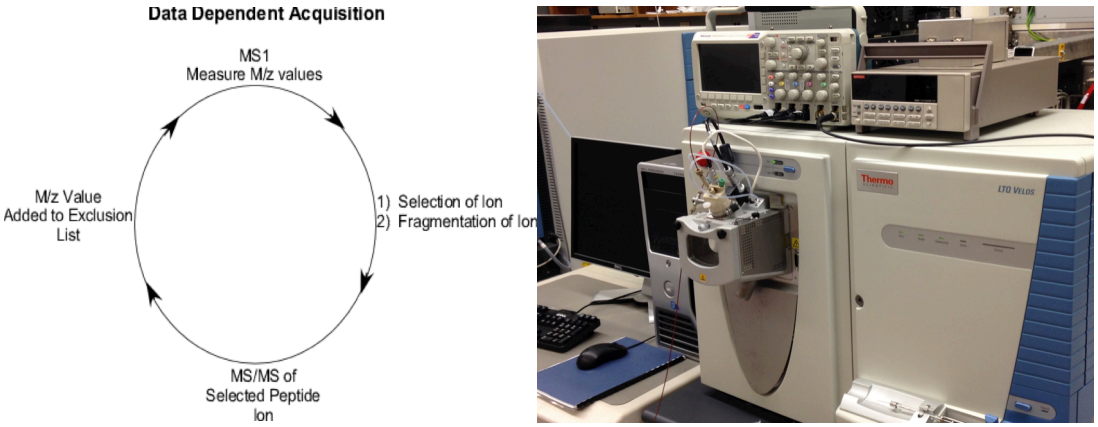


Figure 2. Data dependent acquisition. Computer algorithms on the mass spectrometer record the m/z values of all peptide ions and then based on some criteria selects an m/z value for MS/MS. The m/z value of that ion is then added to an exclusion list so it is not chosen again for some limited amount of time. This capability is present on all modern mass spectrometers.

The powerful proteomic analytical platforms available today result from a co-evolution of HPLC and mass spectrometry technology with the software tools necessary to interpret the data. Tandem mass spectrometry generates large amounts of data, so a key technical hurdle in shotgun proteomics was data interpretation, which has primarily been addressed through the development of algorithmic methods to match tandem mass spectra of peptides to amino acid sequences contained in the protein sequence databases created through genome sequencing and annotation efforts (Figure 3) (7).

This algorithm was extended to identify covalent peptide modifications not available in sequence databases and to search nucleotide sequence databases directly (8). Incorporation of stable isotope labeling into workflows to create relative standards to quantify peptides and proteins necessitated the creation of software tools to interpret the relative ion intensity signals between the heavy and light isotope labeled peptides for large-scale quantitative data (9-12).

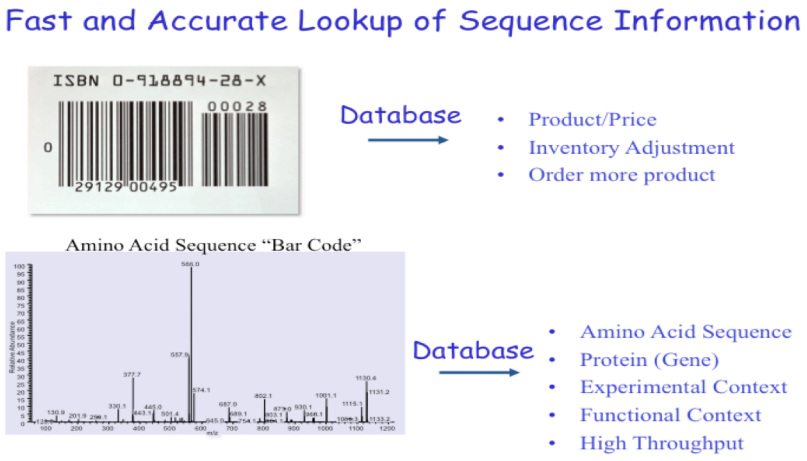


Figure 3. Sequence databases create a resource of protein sequences that can be used to interpret tandem mass spectra of peptides and proteins.

Proteomic software tools include signal processing techniques, algorithmic methods for fast searching, mathematical techniques for spectral comparison, and statistical methods to evaluate results. These tools are combined into pipelines to process tandem mass spectrometry data for large-scale proteomics experiments (Figure 4).

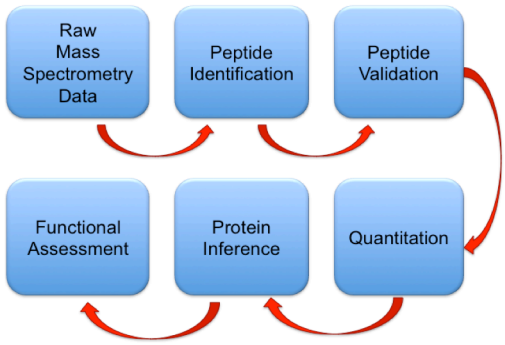


Figure 4. A typical computational pipeline for analysis of mass spectrometry data.

As mass spectrometry technologies and experiments evolve, there are increased opportunities for algorithms and informatics to drive new proteomic capabilities. It was also noted that protein molecules are so rich with detail and their analysis so context dependent, that computational support of diverse workflows is an area in particular need of development for the nation to better leverage and capture the value of this research into complex systems.

A further revolution in mass spectrometry has been the development of methods to analyze intact proteins and their multi-protein assemblies in a process referred to as "top-down" mass spectrometry (11-13). Driven by rapid improvements to mass spectrometers, such as the new 21T FTMS system for high performance mass spectrometry, top down analyses have evolved quickly over the last 10 years (Figure 5).

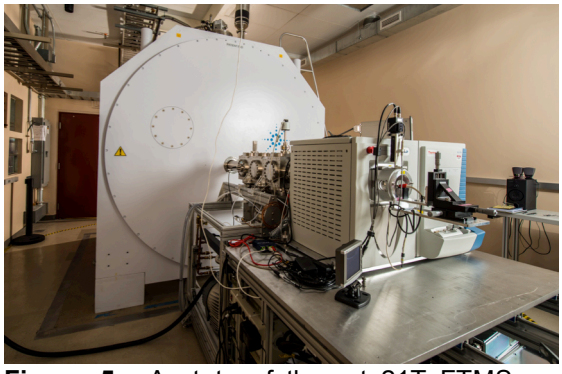


Figure 5. A state of the art 21T FTMS mass spectrometer for high performance top down mass spectrometry.

In the top down strategy, individual protein ions are isolated from complex mixtures of proteins and fragmented to produce collections of fragment ions that provide partial to complete protein sequence information (Figure 6). This top-down protein analysis fragments backbone bonds along the intact protein, capturing co-occurring sequence variations and post-translational modifications across the entire length of the protein, combinations of which form the

proteoforms expressed by a given system (14). Analysis of top-down mass spectrometry data is computationally intensive and thus algorithms to speed accurate interpretation of top down spectra have significant growth potential (15).

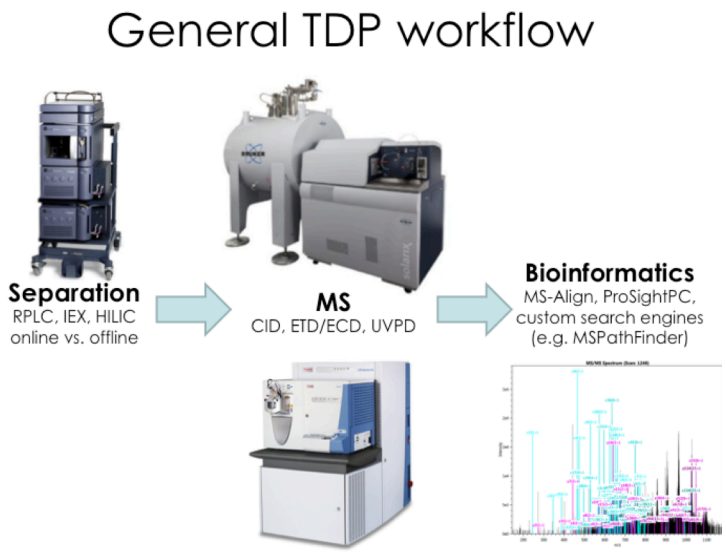


Figure 6. A general workflow for top down mass spectrometry. As the data is very complicated given the size of the proteins analyzed, algorithms to interpret the data are very important.

Biological benefits of proteomics. The combination of mass spectrometry and database searching methods has revolutionized protein biochemistry, disrupting long-standing methods such as Edman degradation and the concept of "one protein, one analysis". This fundamental change in protein analysis has been broadly applied to solve very important and challenging biochemical problems that simply were not tractable using prior methods. The excitement is that proteomics can lead to unexpected biological connections, alter dogma and to provide transformative concepts.

Protein complexes. To carry out their physiological functions, proteins form complexes, creating higher-order structures within cells that define the cellular function via elements such as transmitting signals and controlling enzyme reactions. Determining the composition of these complexes can help determine the proteins involved in specific physiological pathways and functions. Large-scale studies using shotgun proteomic methods to identify the components of complexes have been transformational. The first such application, by McCormack et al., identified protein-protein interactions enriched by immunoprecipitation, affinity enrichment, and co-purification (16). Link et al applied the method to a very large protein complex (5). Later studies examined many complexes from model organisms such as yeast and *Drosophila melanogaster* (17-21). Huttlin et al showed the power of this approach to better understand mammalian biology through the identification of 10,000 human protein-protein interactions (22). Further studies have shown the interactions between mammalian host proteins and viral proteins to illustrate how viruses hijack cellular machinery for their benefit (23-29) providing unexpected generalizable therapeutic interventions for fighting HIV to the common cold. These methods are defining the inner workings of cells at the protein level.

Organelles. As organisms become more complicated, they form more compartments within cells to segregate cellular activities (30-32). For example, the mitochondria are the powerhouses of the cell, generating most of its energy. The mitochondria have their own small genome, but they also contain many proteins derived from the nuclear genome. Proteomic techniques have been instrumental in determining the identities and quantities of proteins in mitochondria from different cell types. Other types of organelles such as peroxisomes, phagosomes, Golgi, processing (P)-bodies, and endoplasmic reticulum (ER) have been studied using proteomic methods to illuminate the functions occurring within these cellular compartments. Subcellular compartments take center stage in specific diseases, such as mitochondrial disease, which results from disease inducing mutations in the mitochondrial genome. In heart disease, the reestablishment of the ratio between the proteins comprising the cardiac myofilament subproteome using a novel pace making program could change how hundred of thousands of patients with heart failure are treated. By understanding how the proteins present in healthy compartments compare with those in unhealthy cells, we can better understand how to treat these diseases.

Cells. Humans have many cell types, which have different functions and express different proteins. The ability to compare protein expression between normal cells and diseased cells provides a means to determine the mechanism of action of disease. An exciting area of rapidly accelerating research is the transformation of skin cells of patients with a disease into pluripotent stem cells which are differentiated into the disease-associated cell type. For example, Brennard et al converted skin cells from schizophrenia patients into neurons in order to compare protein expression with that of normal neurons (33). The ability to study the cell biology of cells with specific disease-associated genetic markers has allowed the discovery of molecular mechanisms underpinning phenotypes associated with both health and disease. Another emerging area of promise is the ability of proteomics to provide a comprehensive catalogue of proteins on the surface of cells, thus helping to define cell types with better precision (34).

Tissues. Tissues are organized communities of cells that form structures to perform a variety of functions. By creating catalogs of the molecular architecture of tissues, we are better able to understand what goes wrong in disease. For example, proteomics provides a powerful tool to study diseases of the brain, such as schizophrenia and Alzheimer's disease, which are difficult to study in cell cultures because the dysfunction stems from disrupted interactions between neurons. McClatchy et al. measured protein expression in mitochondria and synaptosomes from rat brain as a function of developmental time point and brain region showing how protein expression changes over time in different regions of the brain (35, 36). Some affective disorders are thought to be perturbations to

normal development, necessitating an understanding of molecular changes in addition to morphological changes to identify the origin of disease. Proteomic analysis of tumor tissues has revealed differences in protein expression from normal tissue, prioritizing candidate tumor driver genes (37).

Post-translational protein modification. The ability to use mass spectrometry and software tools to identify post-translational modifications (PTMs) has revolutionized our understanding of their regulatory roles in biology. Application of mass spectrometry along with approaches for PTM enrichment and site-specific quantification has been and will continue to be transformative. For example, Huttlin et al. measured the phosphoproteome in mouse, revealing a dramatic contrast in phosphoprotein and phosphorylation sites in 10 different tissues (38). McClatchy et al. determined that signaling pathways for long-term potentiation were disrupted in a rat model for schizophrenia (39). These large-scale methods have allowed studies of ubiquitin, phosphorylation, methylation, O-GlcNAc, acetylation, and other PTMs (40). Mass spectrometry methods which ease the detection and quantification of known PTM can be transformative, especially when they can be used by a broad community (e.g. OGlcNAc) or when methods are developed for under studied PTMs (e.g. Arginylation), allowing new and expanded roles in biology to be uncovered. Thus, for the first time, such studies allow us to understand the connections and cross talk between such modifications in a global manner.

New Opportunities for Proteomics: Next-Generation Proteomics.

Biological systems contain four levels of protein information (Figure 7). The first level is the molecules present in the cell. Defining that information is no small feat, given the many cell types that can change dynamically. The most mature technology to determine what proteins are present is bottom-up proteomics. Efforts to push the technology to measure entire proteomes are nearing completion. However, these studies don't always capture the isoforms of proteins that are present, the complete inventory of modifications, or the patterns of modifications with regulatory roles. Capturing this information will require increasing bottom-up sequence coverage to near 100% for all proteins present. To assign functions to all protein proteoforms we must first establish technologies to measure them well. This task may be impossible using bottom-up data only, so we must develop robust, high-throughput methods to identify intact proteins. Technological innovations are needed to create nearly complete fragmentation of the protein backbone, so all amino acid residues and modifications can be defined within a protein's sequence. Additionally, we need robust quantitation of protein amounts to correlate with other key molecules such as mRNA, miRNA, and metabolites. These goals present significant technological and computational challenges.

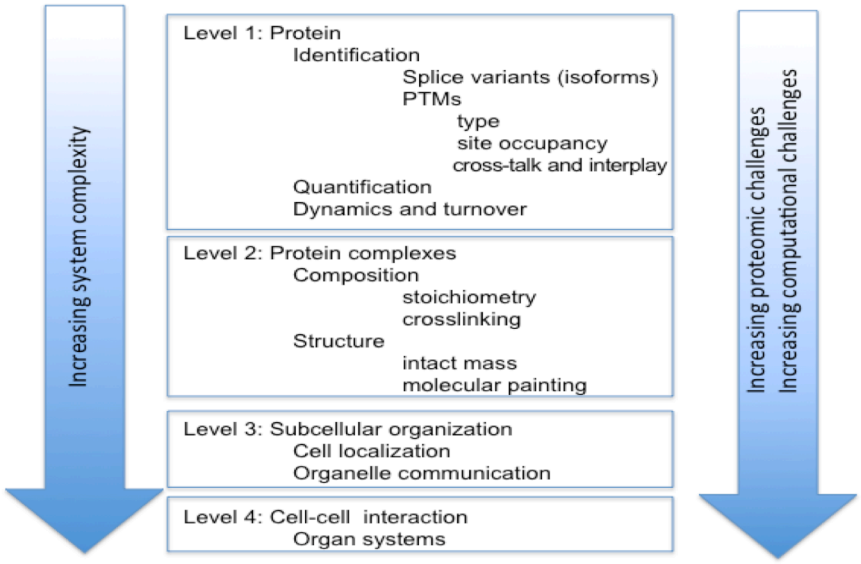


Figure 7. The four levels of biological information that can be mined from cells.

The second level of information is the higher-order structure within cells. Proteins exist primarily as complexes with other proteins, which form the operational units. Defining the higher-order structure of protein assemblies within cells reveals their organization within the cell and functional relationships between proteins. In short, these data are highly valuable because they are obtained at a different level in the naturally hierarchical organization of living things. This is where working strategically and as community, major strides will be realized in the decade ahead.

While in the long-term a cell may change the expression level of a protein, for rapid response to stimuli protein activity is regulated through the transient addition of post-translational modifications that change its interaction partners. Traditional methods, such as immunoprecipitation of proteins under non-denaturing conditions, can determine protein–protein interactions, but are cumbersome and slow for studying the dynamics of large numbers of complexes. New strategies using covalent protein cross-linking to determine interactions are in development, but technical challenges remain, such as the dynamic range of labeling, recovery of cross-linked peptides, and robust software tools to identify the cross-linked peptides. Additionally, methods such as hydrogen/deuterium exchange (HDX), fast photochemical oxidation of proteins (FPOP), and Stability of Proteins from Rates of Oxidation (SPROX) are used to define surfaces of interactions between proteins and protein and small molecule ligands (41-46). A future goal for proteomics would be to define all protein surfaces *in vivo* as a means to elucidate how proteins are folded *in vivo*, when they misfold in disease and how protein complexes are arranged in the cell. Increasingly protein complexes are being studied under physiological conditions using native mass spectrometry and computational tools, which may help to increase the resolution of these methods and their ability to define the structure of complexes.

The third level of information is how proteins and protein complexes are organized or localized within the entire cell. Cell imaging methods work very well and with more success large 2- and 3-dimensional datasets are now available. A major current bottleneck in mass spectral imaging is identifying the proteins and proteoforms detected during analysis. In contrast, mass spectrometry can identify previously unknown proteins inside cellular compartments. As methods to enrich subcellular compartments are imperfect, the development of new mass spectrometry informatics strategies to decipher localization such as protein correlation profiling will provide powerful tools to understand the role of cellular compartments in disease (47-49).

The fourth level of protein information is systems organization and the higher-order structure information also includes how different cells are organized within tissues, a topic that is particularly relevant to the organization of neurons in the brain and how the brain interacts with the other organs.

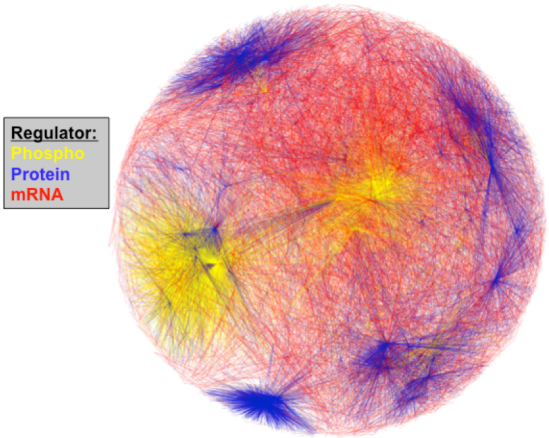


Figure 8. The connectivity of regulatory networks is greater if regulators are measured as proteins.

As we make inroads into interrogating each of these four levels of protein information, new frontiers will emerge on how to connect them together and realize a protein molecular view of cellular physiology that is at once specific to individual protein pools as well as dynamic with regard to their space- and time- evolution. This will require a new cadre of colleagues and developments in data science to aggregate diverse information, rendering them integrated to describe a biologically functional proteome in all four dimensions. When we gain a holistic portrait of protein pathways and networks on a large scale, we shall be able to define the behavior of protein subpopulations in detail (e.g., only the phosphorylated protein molecules); we shall understand with whom they interact, at what locale (e.g., in a transient membrane protein complex) and for how long that molecular event will last (i.e., half-life of the complex during disease development) (Figure 8). With these rich descriptions, we will be able to better link proteomic characteristics to phenotypic outcomes, and eventually automate the process of discovery by coupling spatio-temporal protein quantitation to functional characterization.

What has become clear from 20 years of proteomics and mass spectrometry research is the synergistic relationship between instrument and software development. When software enables the creation and interpretation of larger-scale datasets, new features are added to instruments to take advantage of these capabilities. Alternatively, improvements in technology also drive new software development. For this reason, it is important to consider the future of software in the context of current and future instrument capabilities, and to create operational efficiencies that better democratize access to proteomics technology. The partnership between mass spectrometry technology (data generation) and computational firepower (data processing) has become increasingly interdisciplinary and has led to formulation of visionary approaches to high-impact biological problems. Software and algorithms have long driven innovations in mass spectrometry, and this should continue into the future (50).

I. MASS SPECTROMETRY AND DATA GENERATION

Technology development in mass spectrometry is a vibrant and fast moving area. Future developments in capability will likely create improvements in four areas: 1) higher quality (resolution and mass accuracy), more complete, more comprehensive data collection using systematic data acquisition strategies; 2) higher quality, more complete, more comprehensive fragmentation of intact proteins and multiplexing of data acquisition; 3) methods to determine shape, folding, structures of intact proteins and complexes by using a number of auxiliary methods in conjunction with mass analysis, including cross-linking, ion mobility, hydrogen-deuterium exchange, ion-ion chemistry and various MS/MS methods; 4) pan-omic strategies to multiplex the acquisition of data from many different kinds of molecules simultaneously.

Although many of the methods that will enable new types of experiments and types of data to be collected are not yet well defined nor even invented, software tools will surely be key to driving these capabilities and enabling routine or large-scale application. For at least the last 20 years, a synergistic relationship has existed between mass spectrometry and computational data analysis. The availability of new instrumentation and methods yielding new data types or means to acquire data has driven novel software methods, and conversely computation has enabled deeper and more automated analysis of the diverse data.

Technology and data creation. The development of new technologies will likely drive the collection of new types or scales of mass spectrometry data. Central to the structural analysis of ions are methods to create those ions in a robust and durable manner. Native spray methods, which allow ionization of molecules with retention of native conformations, are key to deciphering the higher-order structure of proteins and complexes. Soft ionization techniques, such as Matrix-Assisted Laser Desorption Ionization (MALDI) with Electrospray Ionization (ESI) or Desorption Electrospray Ionization (DESI), present new methods for the study of biological systems. More importantly, methods to dissociate ions in the gas phase (Collision Induced Dissociation (CID), Higher-Energy Collision Dissociation (HCD), Electron Transfer Dissociation (ETD), Surface Induced Dissociation (SID), Infrared Multi-Photon Dissociation (IRMPD), Ultraviolet Photodissociation (UVPD)) or emerging methods such as ion-ion reactions will create more efficient and predictable ways to fragment ions, and thus will improve the ability of computer algorithms to interpret fragmentation patterns. New breeds of smart mass analyzers combining devices such as ion mobility spectrometers will facilitate improved gas-phase ion separations with a wide range of utility for complex mixture analysis and create new computational

opportunities. New ion mobility methods also offer significant potential to broaden and deepen the information extracted from the most complex mixtures; as ion mobility methods are more widely integrated with mass spectrometer platforms, the development of more powerful and faster data analysis algorithms will be critical. Even age-old problems like differentiation of isomers stand to gain traction via new combinations of ion activation methods combined with computationally intensive pattern recognition algorithms. Studies aimed at spatiotemporal characterization of intact molecules (metabolites, peptides and proteins) in complex samples originating from natural or engineered biological systems (at higher spatial resolution then currently attainable) are needed to address spatial relationships and molecular heterogeneity.

Smart MS/MS and dynamic range issues. Software issues also extend to the acquisition of data as a means to increase data and create different data types. Computers operating mass spectrometers continue to grow more powerful according to Moore's law; consequently, intelligent data acquisition strategies become more feasible. For example, real-time strategies could encompass on-the-fly protein expression profiling where only peptides/proteins that are differentially abundant are acquired. An integrated computational approach combining real-time database searching, spectral library matching, and de novo sequencing would provide a powerful strategy for increasing the proteome coverage. Strategic use of ion-ion reaction chemistry also offers many opportunities to manipulate ion properties on-the-fly to enhance sensitivity by coalescence of dispersed ion populations or sequencing of large and heavily modified polypeptides by more effective (*de-novo*) fragmentation. For example, if a peptide from a protein is detected in real time, then peptides from that protein would not be analyzed again. Integration of reactive and predictive real-time algorithms will allow the creation of innovative and powerful new data acquisition strategies.

Beyond protein characterization. Although many impressive developments in mass spectrometry have involved proteomics applications, increasingly attention has shifted to the other common types of –omics (lipidomics, metabolomics) and more specialized ones (metaproteomics, metatranscriptomics, glycomics, microbiomics, proteogenomics). What would be truly transformative is a universal approach across all compound classes, hence "pan-omics". Such an ambitious goal will require robust separation strategies, versatile tandem mass spectrometry methods, and highly integrated computational processing. The computational issues of "pan-omics" are intriguing, as algorithms would need to identify the type of molecules present in the spectrum before embarking on interpretation to minimize computational space. "Pan-omics" also offers the opportunity to design algorithms integrating these heterogeneous datasets and providing a better understanding of biological processes.

Mass spectrometry quantification- New strategies for quantitation have greatly increased experimental capability. Isobaric labeling methods such as Tandem Mass Tags (TMT), isobaric Tags for Relative and Absolute Quantitation (iTRAQ) and NeuCode have revolutionized multiplexed analyses and quantitation of peptides. This can include PTM specific tags that also allow quantification at the site (site occupancy) and in some case enrichment of the PTM-modified peptides (e.g. Cys-TMT and iodo-TMT that labels Cys). There is considerable impetus for development of computationally driven strategies that could enable absolute quantitative measurements of peptides and other molecules by mass spectrometry. At the same time, advancing the understanding of the physical processes and chemical properties of ionization together with machine learning methods to support development of computational algorithms to accurately correct ionization biases offer compelling opportunities to improve absolute quantitation. Moreover, the ability to analyze absolute quantities of molecules by mass spectrometry without the use of internal standards would create a paradigm shift in measurement science.

Protein interactions- Experimental determination and measurement of protein interactions plays a key role in generating hypotheses and gaining deeper understanding of biological processes. Protein interaction dynamics, which have traditionally been very difficult to determine in large-scale, are of particular importance. Protein cross-linking, however, is now emerging as a potential high throughput method for identifying interaction partners. Recent work has included both compelling biological studies and exciting technology developments, including the development of new reagents for cross-linking *in vitro*, and importantly *in vivo*, with the ability to separate the populations based on cellular localization. These methods have been used to establish protein-protein interactions, as well as distance constraints, within complexes. Longer-term goals encompass the development of global cross-linking of proteins in cells to more rapidly identify protein complexes, characterize their dynamics, and even illuminating their 3D structures and topology. New cross-linking reagents that are more compatible with mass spectrometry, in terms of allowing enrichment, quantitation, and rapid identification of interaction partners, will be an important goal over the next decade. In addition, the development of smarter cross-linking strategies integrated with both bottom-up as well as top-down methods will allow in-depth characterization of multimeric protein complexes.

Top-Down- Over the last decade top down proteomic methods (i.e. analysis of intact proteins) have evolved significantly, and the expectation is that they will continue to improve dramatically over the next decade. Key goals for the next 10 years will include the development of new methods for activating and

dissociating proteins. UVPD has shown great promise to improve protein sequence coverage (including localization of modifications), which together with other activation techniques such as ETD, SID, and CID will offer great opportunities to improve intact protein fragmentation. New methods to promote gas-phase "enzyme" digestion are needed to break large proteins into smaller fragments, which can be dissociated using current methods such as a HCD, ETD, or UVPD. More efficient fragmentation techniques could be applied to *de novo* sequencing of intact proteins and could facilitate comprehensive characterization of proteoforms (Figure 9) (e.g. isoforms and proteins containing multiple and/or unknown PTMs).

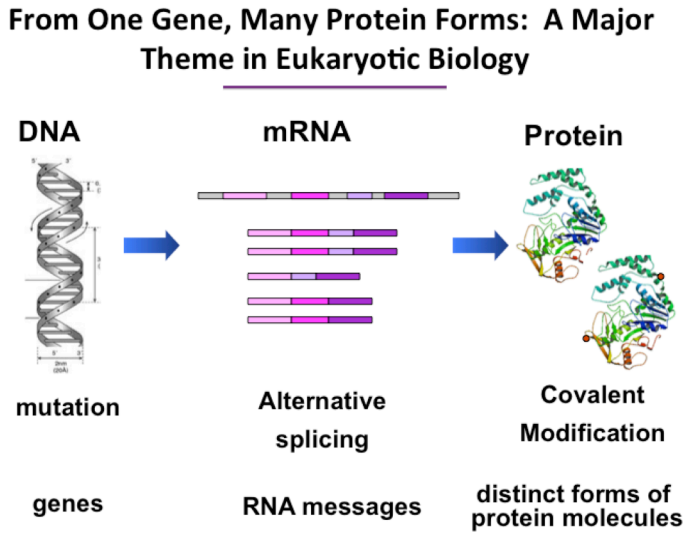


Figure 9. From one gene many different transcripts can be created through alternate splicing. These different gene transcripts can be translated into different proteins that can be modified in different ways. Thus, a single gene can give rise to many proteoforms of a protein.

Imaging mass spectrometry. Imaging mass spectrometry is a rapidly growing field that aims to visualize materials and provide spatial distributions of chemical components in complex samples. Recent advances in imaging mass spectrometry provide significant opportunities for developments in computationally intensive methods, ranging from statistical algorithms that will better facilitate differential comparisons of samples to improved software for construction of more detailed chemical maps to more powerful strategies for deriving quantitative information from spatial profiles. Further improvements in imaging methods will arise from development of better desorption/ionization methods, integrated MS/MS methods to enable direct molecular identification, and to increase the dynamic range of measurements for constituents.

Single-cell analysis. Although the detection limits of mass spectrometers continue to drop for high-throughput analysis, the loftier goal of single-cell proteome analysis or a pan-omics analysis will require an unprecedented level of sensitivity. An even greater emphasis on nanotechnology and microfluidics for sample preparation and separations is essential, such as the development of a smart chip that can accommodate a complex array of sample types and matrix components while integrating multiple

separation modalities. These approaches will require uniting the skills of mass spectrometry specialists with experts in the fields of separations, microfluidics, materials, and biotechnology. Software algorithms as described above would be key to extracting data from these types of analysis. In particular, statistical analysis of single cell data becomes a challenge in the absence of replicates.

Metaproteomics – Genome sequencing methods enable the rapid and thorough sequencing of communities of organisms. Microbiomes are being found to have increasing importance in biology, environmental studies, agriculture, human health and even national security. Humans may have a more symbiotic relationship with our commensal microorganisms than ever before believed. Studies of the microbiomes of the human body have created interesting data sets and present some of the most exciting areas of current biology. The genome sequences of these microbiomes are useful to establish what organisms are present, but they don't reveal the biochemical or metabolic activity of the community, which can reveal synergies or conflicts between organisms. By measuring the metaproteome and metametabolome such activities and interactions can be determined and quantitated, but new databases, algorithms and other software tools capable of dealing with data originating from the large number of organisms present in microbiomes are needed to drive this area of study.

Advances in mass spectrometry technology have been accompanied by both improvements and ongoing challenges in data processing and interpretation. While many of the methods that will enable new types of data to be collected are ambiguously defined or yet to be invented, software tools will be key to driving these capabilities and enabling routine or large-scale application, as discussed in the next section.

II. DATA ANALYSIS – ALGORITHMS AND COMPUTATION

A revolution in protein biochemistry emerged through algorithmic analysis of tandem mass spectrometry data using the sequence information generated by genome sequencing (Figure 10). Over the next decade, powerful new mass spectrometers will evolve as new instrument capabilities and features are invented. The potential of these capabilities to create new data types or to measure new features in biological systems will depend on computer algorithms and software, creating a challenge for substantial improvement to the field of mass spectrometry and proteomics. New opportunities and challenges in mass spectrometry data analysis can be divided into four areas: (1) algorithms, (2) computation, (3) pipelines and workflows, and (4) storage, organization, and interrogation of large data sets.

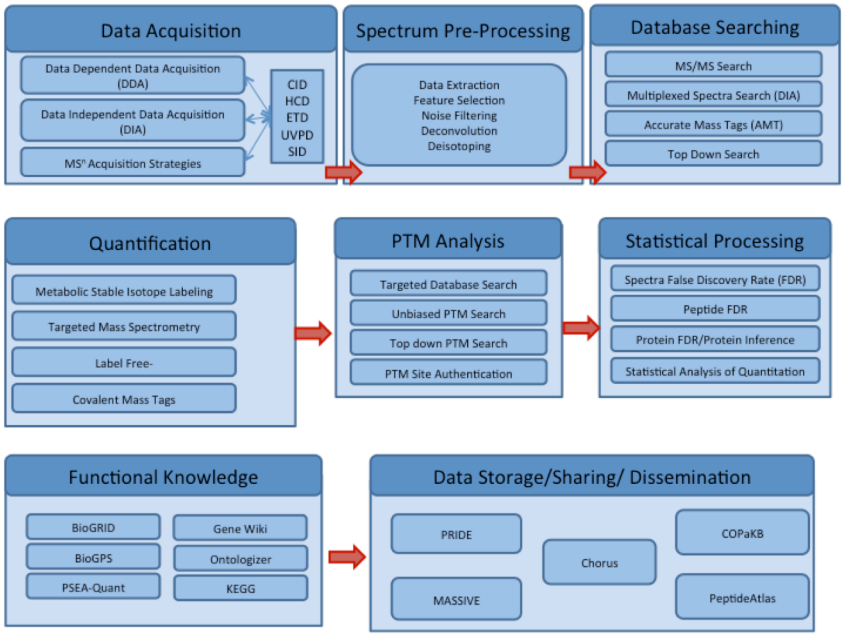


Figure 10. A common workflow for the analysis of large-scale tandem mass spectrometry data.

Algorithms. Improvements in this area would increase accuracy and observation of the peptides in each analysis, ensuring correct proteins and their alternative forms are identified and quantified. Potentially, these improvements will also ensure that new biological SNPs, mutations and PTMs are captured, increasing the biological diversity that can be addressed and maximizing data generation in each sample analyzed on the mass spectrometer.

Algorithms for data acquisition. The first step in mass spectrometry is acquisition of raw m/z data. Mass spectrometers currently carry out real-time computations for charge-state determination, signal-to-noise estimation, and dynamic exclusion for data-dependent precursor ion selection. With the integration of ever more capable on-board processors and advances in real-time algorithms, mass-spectrometers can apply intelligent and adaptive strategies that integrate prior analytical, biological, and

functional knowledge to direct acquisition decisions "on the fly". Some elements of this concept have already emerged, but these ideas can be taken to new levels with an intentional approach. Here, partnership with the private sector could prove valuable to invent and deploy solutions quickly into the field. Improvements in on-board processing capabilities could capture, represent, and model analytical, biological, and functional knowledge for smart decision making in the sub-100 millisecond regime by embedding sophisticated statistical, predictive data-modeling, and machine-learning techniques. Intelligent real time data acquisition could be used to address critical bottlenecks (e.g., dynamic range, scan rates) and increase both value and ease for non-expert, but practitioner laboratories. Improvements in this area would increase accuracy and extent of spectra acquired to maximize proteome coverage in each sample - gathering as much data as possible in each acquisition.

Algorithms for data extraction and pre-processing. An underestimated aspect of the data analysis process is the extraction and processing of spectral data from the manufacturers' proprietary data formats. Extraction of the spectral data to open data-formats facilitates downstream data-analysis by removing operating system, computational platform, and programming language constraints on software. Tools for spectral data extraction also provide an opportunity for semantic enrichment by post-processing the extracted data, to detect and integrate peaks, for semantic data-compression, to separate overlapping ion series, improve post acquisition calibration, identify monoisotopic ions, make use of isotopic clusters, and deconvolute mixed tandem mass spectra. As new technologies that change the data-acquisition paradigm are introduced, semantic data-transformation to computationally reconstruct traditional spectral data-types will be needed to leverage the existing infrastructure of tools, while specialized algorithmic strategies are developed. New mass spectrometers with novel ionization, separation, or detection technologies; new chemistries for analyte fractionation and separation; new data-acquisition strategies and analytical workflows; and improved resolution, sensitivity, and speed will all present new opportunities for improvements in spectral data extraction and semantic representation. This is key as it will harvest as much of the data as possible from a single mass spectrometry run.

Algorithms for search and machine learning. The algorithmic framework for the analysis of traditional data-dependent peptide tandem mass spectra by search of databases and spectral libraries is already well established. This stage of the analysis establishes the identity of the molecular ions observed in the spectral data. For some analytes, in particular, peptides, tools for spectral data search are quite mature, while for other analytes, including proteins, cross-linked peptides, glycopeptides, glycans, lipids, and small molecules, algorithmic improvements and robust and accurate tools are

sorely needed. For many of these analytes, comprehensive databases and spectral libraries are lacking, while for peptides and proteins, sequence databases fail to adequately describe the space of possible proteoforms. As the speed, sensitivity and acquisition capabilities of mass spectrometers and sample-handling automation improve, even the mature algorithmic frameworks for peptide identification may need to be reconsidered. For example, the database-search paradigm set in place 25 years ago for shotgun proteomics must be adapted or augmented to handle multiple fragmentation modes and data-independent acquisition (DIA) methods. We have already seen an explosion in the size of the spectral data from bottom-up proteomics studies, and going forward, we anticipate spectral data search will become a significant computational bottleneck requiring new algorithmic approaches and computational strategies to keep pace with data acquisition. Search algorithms which achieve (sub-)linear search times in the size of the database and spectral data; which optimize the quality and quantity of identifications given a fixed time and memory allocation; and which are suitable for unreliable, heterogeneous, distributed compute resources and a variety of operating systems and computing environments (cluster, grid, cloud) are needed to meet this big-data challenge.

Processes for routine searching of databases and spectral libraries are already well established. For some types of experiments, such as top-down proteomics, optimization of speed, sensitivity, and specificity are necessary to increase scale and throughput of experiments. The development of new algorithms to identify molecules from tandem mass spectrometry experiments or to interpret new types of molecular information through the exploitation of new computational approaches based on machine learning will create new data analysis opportunities. Newer data types or new experimental paradigms that involve data-independent acquisition of mass spectrometry data, imaging mass spectrometry, real-time data analysis, and pan-omic data analysis represent exciting informatics opportunities and challenges. Additionally, multiplexing biological analyses for quantitative top down proteomics (e.g. using metabolic or post-growth strategies with proper computational support) would create significant efficiencies. Furthermore, computational methods could be created to do what cannot be easily done now. For example, enriching organelles to homogeneity is very difficult to achieve, so computational strategies to derive enriched statistically significant signal from noisy data would benefit localization studies. Similarly, new tools could allow researchers to assess protein-protein interactions without having to pull down each complex using methods such as co-elution from chromatographic or electrophoretic separations. Search methods to identify different kinds of molecules from the same analysis would enable pan-omic studies. The search process is fundamental to the analysis of mass spectral data, but inventive methods to increase capabilities to uncover new features or discover new biology should be encouraged.

Characterization of PTMs, Isoforms and SNPs. Beyond automatic spectrum assignment, algorithmic development is needed for analytes without comprehensive databases or spectral libraries, for combinatorially assembled analytes that cannot be readily cataloged or enumerated, and for more detailed, biologically focused characterization of analytes than is possible using spectral data search. The ability to accurately de novo sequence peptides and proteins will create important tools for areas such as metaproteomics to supplement metagenomics information. Better algorithms are needed to identify and localize PTMs in an unbiased fashion especially when peptides contain 2 or more modifications or sequence variations, and algorithms to match detected peptide-level PTMs onto protein-level modification isoforms. Algorithms are needed to characterize complex, biologically important post-translational protein modifications such as N- or C-terminal cleavage and proteolysis, N- and O-glycosylation, and branched multi-protein forms due to ubiquitination or sumoylation. Algorithms coupled to new experimental paradigms to collect and analyze data in 4 D – space, time, quantity, and activity would greatly expand our understanding of biology and create efficient experimental paradigms. New strategies to simplify quantification, especially in complicated systems such as microbiomes, will be important to understand the biology of these complex communities.

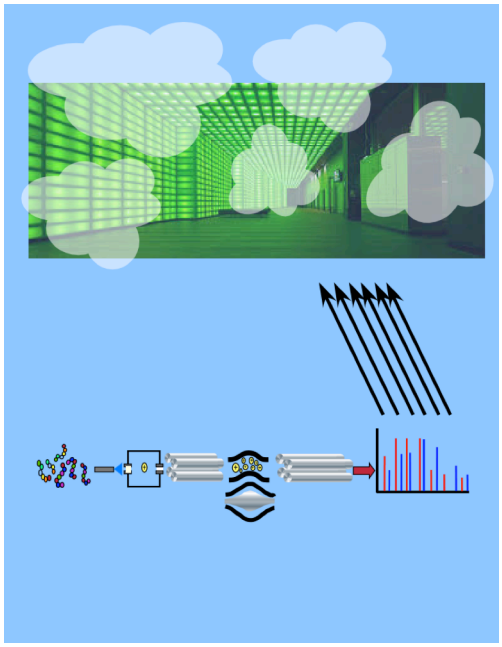


Figure 11. Computing in the cloud will democratize access to high performance computing power.

Computation Requirements. The last 20 years have seen great strides in the democratization of computing power. No longer is access to a supercomputer required for high performance computing. The routine availability of high-performance computing and high-bandwidth networking has enabled computationally intensive methods to transform information retrieval, speech recognition, natural language processing, physical simulation, computer vision, and robotics. Additionally, the increasing availability of computing clusters, clouds, grids and GPUs will provide even greater access to high performance computing as the cost structures of computing environments such as cloud computing improve. While some efforts to employ such computing environments in proteomics have been proposed, more work is needed to explore cost efficiencies, practicalities, scalability and advantages of these environments. A

concerted effort to integrate with centralized scientific computing resources, including the national cyber infrastructure, should be made. As proteomic data sets and genomic datasets (e.g. metagenomic databases) increase in size and as more information is mined from these data sets (such as PTMs, sequence variations, unknown features, etc.) computational power, scalable proteomics platforms and the algorithms to take advantage of that power will be critical. Democratizing computing power for the proteomics community to broadly enable the research will be key to driving future discoveries (Figure 11). Adoption of scalable engineering and big-data techniques which expect and accommodate unreliable computing and failures of data-integrity, and which apply coarsely decomposable processing strategies that minimize inter-processor communication will be important. Keeping individual processing requirements (data-transfer, memory, I/O) modest and moving to the model in which many small (virtual) computing resources can be applied, will help to minimize the need for large monolithic computers. Where possible the development of open-source tools should be encouraged. This will avoid per-processor software licensing charges, which are an additional expense for (virtual) computing resources. To ensure the development of the most effective tools, the creation of teams of software developers and users will be necessary with encouragement to disseminate validated tools efficiently using both open-source and software-as-service models. Here development of tools can be clearly differentiated from those validated and ready for robust, high-value use by the community for reproducible findings in publications. Such a working model will help a diverse community leverage work from other contributors to build more advanced and scalable tools.

Pipelines, Workflows, Laboratory Information Management Systems (LIMS). Increasingly, software tools for proteomics are organized into pipelines, workflows, or LIMS to better capture experiment design, to promote reproducibility of complex data analysis workflows, and to manage large datasets and complicated experiments. Pipelines should handle the complexities of managing jobs in large-scale, unreliable, heterogeneous computing environments, managing on-demand compute resource scaling, data-transfer, job scheduling and failure, and synchronization as needed. These pipelines should be modular, with interchangeable software tools supporting well-established input-output file formats, preferably focusing on standard open formats already developed by the proteomics community. These pipelines should support a variety of mass spectrometers, open spectra data formats, spectra data search tools, and statistical-significance estimation techniques, to avoid platform lock-in and lack of cross-platform reproducibility, except where significant advantage is available. Pipelines should record the provenance of its results, including software, input files, and parameters used at each step. Software libraries for a variety of programming languages and platforms, designed to support the implementation of post analyte-identification biological and functional inference, are

needed to accommodate the vast array of studies capable with mass spectrometry-based proteomics and to have sufficient scalability to allow large-scale studies. To better encourage data and metadata depositions into repositories, direct interfaces between these pipelines and public repositories will be essential.

Local and Public Storage of Large Data Sets. Mass spectrometry data sets alone and collectively capture a considerable amount of analytical, biological, and functional information, whether or not researchers have the computational tools to extract this information. This information exists at many levels, from performance metrics of a specific mass-spectrometer, to the characteristics of peptides and proteins observable in a specific analytical workflow, to the abundance of various proteins in specific types of cells, and the changes in protein isoforms in contrasting phenotypic samples. Here too, we have a semantic representation challenge, as we strive to expose the characteristics of each spectral data set most useful for deriving analytical, biological, and functional knowledge. As data sets increase in size, the ability to analyze and mine these datasets will create an enormous opportunity to add value to biological studies and data. To make such analyses possible, tools for individual laboratories to use, explore, and mine large scale data sets, and to compare the results to those of other groups, are needed. Such tools will encourage the collection, retention, and reutilization of data and facilitate the sharing of data via public repositories. The creation of flexible and smart data repositories can facilitate spectral re-utilization as a library and should support the progression towards community wide consensus interpretations to translate data into a reusable knowledge base resource. Crucial for this effort is the capture of appropriate experimental design and analytical workflow metadata – without this metadata, the utility of such repositories is limited to analytical knowledge. Repositories should store spectra in open data formats in addition to the manufacturer's proprietary data-formats, so that users are not required to use vendor software to access the spectra. Repositories of derived analyte and biological information, such as identified peptides, characterized PTMs, and inferred proteins linked to spectral data repositories, can be used for semantics-based queries and for analytical, biological, and functional knowledge inference impossible from uninterpreted spectral data. These repositories will be more technology- and technique-agnostic, avoiding “data rot” as technologies, chemistries, and techniques change. Software tools to crowd source, wiki source or game source the interpretation of spectra that are intractable to current software-based interpretations can aid in the creation of new data and create communities of spectral “annotators” (Figure 12). Spectra can also serve as physical evidence supporting database information on protein sequence features. For example, there is increasing evidence that small peptide ORFs not easily identified in the genome sequence by

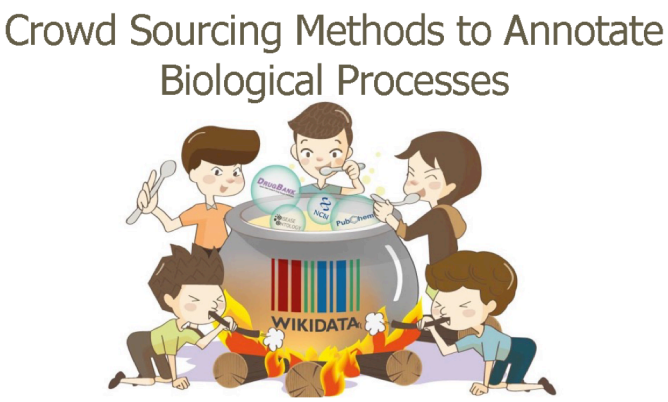


Figure 12. An increasingly attractive strategy to contribute knowledge and interpret information is through the use of the "crowd sourcing" methods as developed through the Wikipedia model.

informatics are real entities with as of yet obscure biological functions. Spectral resources may help to identify more examples of these cases, assisting in determining their biological roles. Of particular importance in the context of all the algorithmic challenges mentioned above, smart data repositories should interact with the community to develop `reference datasets' that are especially suited (and maybe even specifically designed) to drive and benchmark software developments aimed at meeting specific computational needs. A grand challenge and opportunity is creation of a centralized proteomics resource that aggregates knowledge and allows for asking intuitive biological questions, a portal that would allow biologists to seamlessly travel through many data levels. To better enable the broader biological community to use mass spectrometry data, educational resources need to be created such as tutorials, books, MOOCs, videos, and workshops at non-traditional conferences. If mass spectrometry/proteomics data could be made as useful to the broader scientific community as genomics data has become that would be a remarkable achievement.

III. TRANSLATING DATA TO KNOWLEDGE

A critical step in all scientific inquiries is deriving knowledge from data, thus creating a fuller understanding of the underlying mechanisms and principles. Proteomics experiments typically involve a number of parallel data acquisitions, which may involve phenotypically contrasting samples or different growth or disease conditions, as well as biological and technical replicates required by the experimental design, and separation, fractionation, and enrichment required by the experimental methods. The analytes identified from each acquisition must be merged, in a way that depends upon the experimental design, to obtain a comprehensive and coherent picture of the experimental results. Even the most basic steps in this process, such as protein inference from peptide identifications, raise nontrivial issues of prior knowledge, Bayesian inference, multiple hypothesis testing, and FDR (false discovery rate) estimation. "Second-order" inference, for example, principled inference of sequence or splice variants or non-transcriptional regulation, require methods yet to be developed. Where experimental designs incorporate phenotypically contrasting samples, quantitative biological inference may be projected forward, based upon well-characterized pathways, to make functional inferences – here too, new tools and inference techniques are required. As data sets have grown in size, new paradigms for assessment of data are needed, especially for new data types such as DIA "libraries", top down protein data, cross-linking data, and metaproteomic data.

Four focus areas stand out in the effort to develop data-to-knowledge tools: (1) metadata, meaning complete and accurate machine-readable descriptions of the experimental conditions and parameters, (2) meta-analysis of related data sets to derive inferences about biological processes or diseases, (3) integration of diverse data types (omics data), and (4) outreach to the community to crowd-source the analysis of data, and to teach scientists and other stakeholders about mass spectrometry and proteomics data. A unifying vision for the effort would be an intuitive, user-friendly, centralized portal that enables discovery, exploration, and validation of biological hypotheses.

Metadata. The value of proteomics data repositories critically depends upon the quality and accessibility of the metadata accompanying the mass spectrometry measurements. Metadata includes organisms and strains, growth and collection conditions, sample preparation methods, and mass spectrometry data acquisition strategy. Current data repositories rely on depositors to describe their data accurately and choose appropriate keywords for search and retrieval.

The field needs improved methods for both data deposition and retrieval. On the deposition side, controlled vocabularies, clearly defined terms, "persuasive technology" questionnaires, and crowd-

sourcing could improve the quality and compliance of meta-data. Most data acquisition metadata could be extracted automatically from method settings and scan headers. Running fast preliminary searches to look for abundant proteins and modifications could check metadata, such as organism and sample preparation methods, and depositors could be alerted if these searches show errors or inconsistencies in the metadata. Preliminary searches could also estimate quality metrics such as chromatographic peak width, digestion specificity, signal-to-noise ratios, and m/z calibration. Depositors, users, or bots could curate spectra for spectral libraries, and connect related data sets based on organism, tissue, disease, or protein content. Data repositories should maintain audit trails to ensure data integrity.

On the retrieval side, improved methods could employ whole-text indexing, natural language processing, automatic summarization, recommendation and reputation systems, and other technologies developed for Internet search. Methods could include text-based retrieval, along with “find similar data sets” and “recommended for you” functionality.

Sustained government support will be required for data storage, as well as licensing agreements, privacy protection, standardized formats, format conversion, replication, data compression, error-correcting codes, and “cold storage” to assure that older data is not lost and that the most popular new data can be accessed quickly and reliably.

Meta-analysis. Current practice analyzes proteomics data sets to compile lists at various levels of detail (spectrum assignments, peptides, and proteins), along with relative or absolute quantitation of the detected species. False discovery rate is estimated and controlled by a target-decoy strategy that models false spectrum assignments using fictitious sequences, for example, reversed protein sequences. These methods measure a single type of error, namely false peptide sequences, but do not measure other types of errors, such as incorrect proteins, misplaced PTMs, and so forth. Typically data is analyzed to the level required for publication by those who produced the data. Mass spectrometry data sets are rarely re-analyzed and mined for new results after the initial publication; rarely analyzed to higher levels of organization, such as complexes or co-regulation; and rarely correlated with related data sets, for example, homologous systems in other species. Multiple accessions of single proteins and single accessions of multiple proteins make it difficult to correlate results of searches against slightly different protein databases.

Research should advance on multiple fronts. Improved statistical methods could estimate false discovery rate at peptide, modification, and protein levels. The notion of “false discovery” could be

further developed to better match downstream inference; for example, a glycopeptide assignment with a correct peptide but incorrect glycan still supports the inference of glycosylation, but a cross-linked or disulfide-bonded peptide pair with one correct and one incorrect peptide leads only to false structural inference. More flexible and extensible data structures for protein sequences could facilitate meta-analysis by handling splice variants and protein families more gracefully. Re-analysis could uncover phenomena, such as sequence variants and PTMs, not considered in the initial publication. Re-analysis assumes special importance for data-independent acquisition methods, because the initial analysis may target only a subset of the peptides and proteins present in the sample. Version control systems, as currently used in software development, could organize the multiple analyses of data sets. The proteomics equivalent of full-text search could determine whether sequence variants or PTMs had been observed in other data sets; indexing at multiple levels (proteins, peptides, PTMs, even peaks in unassigned spectra) would enable fast retrieval. We imagine a "PTM BLAST" tool that, given a PTM site, could determine the most similar PTM sites in large data repositories. Crawlers could extract information from Pubmed and other sources to add references to PTM BLAST search results. Such a tool would enable the discovery of correlated PTMs, novel PTMs, cross-talk, co-regulation, sequence motifs, and pathways. Crowd-sourcing would improve the quality, usability, and interpretation of the analyses.

Omics integration. High-throughput mass-spectrometry-based proteomics would be scarcely possible without genomics. Perhaps not so widely known is that proteomics delivers great benefits to genomics. Proteogenomics has already improved gene calls, determined exact start/stop boundaries, and found novel splice variants for a number of sequenced organisms. Mass spectrometric analysis of intact bioactive peptides, such as hormones, toxins, and neurotransmitters, provides information inaccessible to genomics and transcriptomics, because the activity of these peptides often depends upon post-translational processing. Antibodies and antigens provide other compelling examples of molecules best studied by a combination of proteomics, genomics, and transcriptomics. Mapping complete human antibody repertoires or the full diversity of human leukocyte antigens are "big science" challenges that could lead to medical breakthroughs.

Of course the genomics-to-proteomics flow of information has not ceased with the completion of the reference human genome and the sequencing of all the common model organisms. In fact, Next Generation Sequencing (NGS) can now affordably produce "on-demand" databases for subspecies, strains, microbial colonies, and individuals, and even personalized cell genome databases such as libraries of cancer mutations, gut microbiomes, or B-cell clones. The challenge now falls back onto

proteomics to make use of these databases to obtain more complete coverage of proteins and proteoforms.

Benefits can also be gained from integrating proteomics data sets with the higher functional levels. Proteomics can combine with glycomics, lipidomics, metabolomics, imaging, and pathology to provide molecular views of healthy and diseased tissues. Online resources of mass spectrometry imagery could be a great aid to diagnosis and treatment. Another valuable connection can be made to structural biology, because proteomics can assay protein modifications and variants much more easily and rapidly than X-ray crystallography, which requires a pure sample of the proteoform. MS-based methods for assaying structure, such as hydrogen/deuterium exchange, oxidative footprinting, and cross-linking, have been gaining popularity in recent years; these methods and other methods yet to be developed will become much more common as reagents, methods, and instruments continue to improve. Data visualization will play a key role in omics integration. The recently published draft maps of the human proteome call for proteome browsers, analogous to genome browsers to better examine the details of the data. Linking proteomics data to the Protein Data Bank would enable the visualization of PTMs and cross-links on three-dimensional structures.

Outreach. Biomedical research would be greatly advanced by accessible and understandable proteomics information. Proteomics specialists should make an effort to teach the wider research community what proteomics can and cannot do. The centralized portal we envision should include “Mass Spectrometry 101” to educate prospective users of proteomics and proteomics data.

IV. RESOURCES NEEDED – INFRASTRUCTURE, INSTRUMENTATION, COMPUTATION

Open source software consortia to develop validated tools.

Most software tools are developed by individuals with a data set in hand to test their software. Once completed the software is made available to the community where it may or may not solve the intended problem. This "lone wolf" approach to software development does not necessarily result in tools which are usable or needed by the community unless the developer is willing to invest a lot of time engaged with the community of users. To improve on this model, we advocate a model based on the "wolf pack", a cohesive and collaborative team of programmers and users that interact closely to solve specific informatics problems in a manner that is immediately useful for the field. This wolf pack model will ensure that important problems are solved in a manner that is user friendly and valuable to the end user and stakeholders. Too often open source software is developed and made available to users with instructions to fix the bugs and improve the software because "they have access to the code". Creating opportunities for team efforts in the development of new tools will ensure the right tools are developed and they are well tested before release to the field.

As the development of new methods and techniques does require access to instrumentation, the creation of a path to acquire instrumentation for technology development would help facilitate the process.

The availability of high performance computing for both development and use of computing tools will be essential in an era of big data. Cloud computing is altering the landscape for the availability of high performance cluster based computing. Since most cloud platforms are commercial in nature the question arises as to the suitability of this solution for academic computing needs. The traditional supercomputing center model for access to high performance computing is antiquated for future on-demand needs for high performance computing. Thus an academic model for cloud computing should be explored to assess the financial viability of such a model relative to reliance on commercial cloud computing and a pay-as-you-go model. Given the mandate, the National Cyber infrastructure can be leveraged using existing administrative models to greatly address these challenges.

V. GRAND CHALLENGES AND RECOMMENDATIONS

The workshop explored and discussed challenges for the future of proteomics mass spectrometry and proteomics informatics. Important challenges were identified, ranging from predictive and reactive software for data acquisition to integration of proteomic results with the biological data infrastructure. Thus significant opportunities were identified to move the field forward over the next 10 to 15 years, which should help drive future discoveries in molecular biology and new types of clinical applications.

1) **Software Tools and Algorithms.** Software to interpret data created by mass spectrometers can help drive the development and adoption of new experimental methods. Recent advances in instrument design have created new hybrid mass spectrometers with capabilities for CID, ETD, HCD, and UVPD. These highly effective methods to dissociate ions can be used serially or in combination to create new data acquisition paradigms for more systematic and comprehensive collection of data from peptides, intact proteins and whole protein complexes. For example, multiplexed and multi-modal MS^n methods that collaboratively and cooperatively work with computational methods could be developed. Thus, new opportunities to develop software tools and algorithms for innovative and systematic data acquisition paradigms are possible by integrating the development of new mass spectrometry methods with software development.

2) **Parallelized data Acquisition Strategies.** A fundamental shortcoming of mass spectrometry is the serial nature of data acquisition and sample analysis. Creating parallelized data acquisition strategies is a common method to increase throughput and efficiency in data acquisition. Through a combination of novel data acquisition strategies and the software tools to interpret the data, new strategies to multiplex data acquisition strategies for peptides, intact proteins and whole protein complexes could be possible. Multiplexing samples through the use of isobaric covalent tagging methods has created enormous opportunities for biological analyses. Can other forms of multiplexing through the use of sophisticated statistical analyses, predictive data modeling, and machine-learning techniques be created? Multiplexing data acquisition and sample analysis could greatly improve throughput, efficiency and scale of mass spectrometry experiments. Thus, strategies to create new ways to massively parallelize mass spectrometry analyses could create new experimental economies for proteins biochemistry in much the same way this has been done in Next Gen Genomics.

3) **New Data Formats and Compression Methods.** Mass spectrometry data sets have been increasing at a fast rate and will continue to do so in the future. A general trend over the last decade

has been the gradual increase in instrument scan speeds by ~2-5 times every 2 years. An increasing amount of data and the potential for data sets to contain denser data with the development of new types of experiments will necessitate new data file formats and data compression methods to enable the movement of these data sets among collaborators and storage facilities. Additionally, the larger data sets will require the development of high throughput and large-scale data analysis tools and computing architectures to create durable and accurate interpretations of mass spectrometry data for peptides and intact proteoforms will be essential.

4) **Mass Spectrometry-Based Structural Biology of Endogenously Formed Complexes**. Mass spectrometry is increasingly contributing to structural biology analyses. Most current studies involve *in vitro* studies of proteins, their proteoforms, and multi-protein complexes with and without their bound ligands. A future drive to study the analogous entities formed *in vivo* will help determine their native biological structures and their dynamics over the course of biological processes. These goals will necessitate the development of robust software tools for MS-based structural biology of endogenously-formed complexes, including large-scale cross-linking experiments, residue-level specificity in hydrogen/deuterium exchange, and covalent labeling methods including oxidative footprinting. Furthermore, connecting data from these methods with other types of *in vivo* analyses will increase the demand for computational interpretation and rendering of large and complicated data sets. Thus, the development of methods to create and capture mass spectrometry-based structural information *in vivo* and to combine that data with protein structural prediction algorithms to create high-resolution predictions of native cellular protein structures and protein complexes will drive our understanding of cellular biology and biological mechanisms.

5) **Development of New Statistical Tools For Mass Spectrometry Based Analyses.** As big data emerges in mass spectrometry and proteomics, these larger data sets will present unique opportunities. Many of the challenges of big data are not related as much to the size of the data sets, but as to the existence of noise and errors in the data. Larger data sets will require the development of new statistical methods for False Discovery Rate (FDR) estimation at multiple levels of organization (spectra, peptide, proteins, organisms, higher taxa) and multiple levels of error (incorrect localization, modifications, sequence, protein family). To improve statistical analyses, methods to assess and evaluate the quality of spectra to eliminate poor quality spectra or spectra of noise or non-peptides will help this process.

6) **Integration of Mass Spectrometry Data And Interpretations with Existing Knowledge-Bases.** As the overriding goals of many experiments are to discover new biological insights, the development of tools to integrate mass spectrometry data and interpretations with existing knowledge-bases to speed data to knowledge interpretations. As more biological data are collected they are stored in databases to create knowledge bases. These databases can range from repositories of data to highly sophisticated curated knowledge bases. To speed the analysis and interpretation of proteomic data, these knowledge bases should be affiliated with tools to mine this information. Better affiliation of databases can be accomplished by creating consolidation sites that collect specific types of data from all existing databases or by creating web crawlers that search on demand using specific queries to find all available data on the web.

VI. REFERENCES

1. Hunt DF, Yates JRd, Shabanowitz J, Winston S, Hauer CR. Protein sequencing by tandem mass spectrometry. ***Proceedings Of the National Academy Of Sciences Of the United States Of America***. 1986;**83(**17):6233-7.
2. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. ***Science***. 1989;**246(**4926):64-71. PubMed PMID: 2675315.
3. Stahl DC, Swiderek KM, Davis MT, Lee TD. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. ***JAmSocMass Spectrom***. 1996;**7**:532-40.
4. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR, 3rd. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. ***Anal Chem***. 1997;**69(**4):767-76.
5. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR. Direct analysis of protein complexes using mass spectrometry. ***Nature Biotechnology***. 1999;**17(**7):676-82. PubMed PMID: ISI:000081296900028.
6. Washburn MP, Wolters D, Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. ***Nature Biotechnology***. 2001;**19(**3):242-7. PubMed PMID: ISI:000167283100024.
7. Eng JK, Mccormack AL, Yates JR. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. ***Journal of the American Society for Mass Spectrometry***. 1994;**5(**11):976-89. PubMed PMID: ISI:A1994PP71300004.
8. Yates III JR, Eng JK, McCormack AL. Mining Genomes: Correlating tandem mass spectra of modified and unmodified peptides to nucleotide sequences. ***AnalChem***. 1995;**67**:3202-10.
9. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. Accurate quantitation of protein expression and site-specific phosphorylation. ***Proc Natl Acad Sci U S A***. 1999;**96(**12):6591-6.
10. Paša-Tolic L, Jensen PK, Anderson GA, Lipton MS, Peden KK, Martinovic S, Tolic N, Bruce JE, Smith RD. High throughput proteome-wide precision measurements of protein expression using mass spectrometry. ***JAmChemSoc***. 1999;**121(**7949):7950.
11. Kelleher NL. Top-down proteomics. ***Anal Chem***. 2004;**76(**11):197A-203A. PubMed PMID: 15190879.
12. Ahlf DR, Thomas PM, Kelleher NL. Developing top down proteomics to maximize proteome and sequence coverage from cells and tissues. ***Curr Opin Chem Biol***. 2013;**17(**5):787-94. PubMed PMID: 23988518; PubMed Central PMCID: PMC3878305.
13. Benesch JL, Robinson CV. Mass spectrometry of macromolecular assemblies: preservation and dissociation. ***Curr Opin Struct Biol***. 2006;**16(**2):245-51. PubMed PMID: 16563743.
14. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. ***Nat Methods***. 2013;**10(**3):186-7. PubMed PMID: 23443629.
15. Meng F, Cargile BJ, Miller LM, Forbes AJ, Johnson JR, Kelleher NL. Informatics and multiplexing of intact protein identification in bacteria and the archaea. ***Nat Biotechnol***. 2001;**19(**10):952-7. PubMed PMID: 11581661.
16. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates III JR. Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. ***Analytical Chemistry***. 1997;**69(**4):767-76.
17. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sorensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M.

Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. ***Nature***. 2002;**415(**6868):180-3. PubMed PMID: 11805837.

18. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, Remor M, Hofert C, Schelder M, Brajenovic M, Ruffner H, Merino A, Klein K, Hudak M, Dickson D, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. ***Nature***. 2002;**415(**6868):141-7. PubMed PMID: 11805826.
19. Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A-M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. ***Nature***. 2006;**440(**7084):631-6.
20. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrin-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St Onge P, Ghanny S, Lam MH, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles CJ, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. ***Nature***. 2006;**440(**7084):637-43. PubMed PMID: 16554755.
21. Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O, McKillip E, Shah S, Stapleton M, Wan KH, Yu C, Parsa B, Carlson JW, Chen X, Kapadia B, Vijayraghavan K, Gygi SP, Celniker SE, Obar RA, Artavanis-Tsakonas S. A Protein Complex Network of Drosophila melanogaster. ***Cell***. 2011;**147(**3):690-703. PubMed PMID: 22036573.
22. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, Tam S, Zarraga G, Colby G, Baltier K, Dong R, Guarani V, Vaites LP, Ordureau A, Rad R, Erickson BK, Wuhr M, Chick J, Zhai B, Kolippakkam D, Mintseris J, Obar RA, Harris T, Artavanis-Tsakonas S, Sowa ME, De Camilli P, Paulo JA, Harper JW, Gygi SP. The BioPlex Network: A Systematic Exploration of the Human Interactome. ***Cell***. 2015;**162(**2):425-40. PubMed PMID: 26186194; PubMed Central PMCID: PMC4617211.
23. Weekes MP, Tomasec P, Huttlin EL, Fielding CA, Nusinow D, Stanton RJ, Wang EC, Aicheler R, Murrell I, Wilkinson GW, Lehner PJ, Gygi SP. Quantitative temporal viromics: an approach to investigate host-pathogen interaction. ***Cell***. 2014;**157(**6):1460-72. PubMed PMID: 24906157; PubMed Central PMCID: PMC4048463.
24. Cristea IM, Rozjabek H, Molloy KR, Karki S, White LL, Rice CM, Rout MP, Chait BT, MacDonald MR. Host factors associated with the Sindbis virus RNA-dependent RNA polymerase: role for G3BP1 and G3BP2 in virus replication. ***J Virol***. 2010;**84(**13):6720-32. PubMed PMID: 20392851; PubMed Central PMCID: 2903289.
25. Cristea IM, Moorman NJ, Terhune SS, Cuevas CD, O'Keefe ES, Rout MP, Chait BT, Shenk T. Human cytomegalovirus pUL83 stimulates activity of the viral immediate-early promoter through its interaction with the cellular IFI16 protein. ***J Virol***. 2010;**84(**15):7803-14. PubMed PMID: 20504932; PubMed Central PMCID: 2897612.
26. Shah PS, Wojcechowskyj JA, Eckhardt M, Krogan NJ. Comparative mapping of host-pathogen protein-protein interactions. ***Curr Opin Microbiol***. 2015;**27**:62-8. PubMed PMID: 26275922.
27. Davis ZH, Verschueren E, Jang GM, Kleffman K, Johnson JR, Park J, Von Dollen J, Maher MC, Johnson T, Newton W, Jager S, Shales M, Horner J, Hernandez RD, Krogan NJ, Glaunsinger BA. Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for

late genes. ***Mol Cell***. 2015;**57(**2):349-60. PubMed PMID: 25544563; PubMed Central PMCID: PMC4305015.

28. Naji S, Ambrus G, Cimermancic P, Reyes JR, Johnson JR, Filbrandt R, Huber MD, Vesely P, Krogan NJ, Yates JR, Saphire AC, Gerace L. Host Cell Interactome of HIV-1 Rev Includes RNA Helicases Involved in Multiple Facets of Virus Production. ***Molecular & Cellular Proteomics***. 2012;**11(**4). PubMed PMID: ISI:000302786500021.
29. Jager S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, Shales M, Mercenne G, Pache L, Li K, Hernandez H, Jang GM, Roth SL, Akiva E, Marlett J, Stephens M, D/'Orso I, Fernandes J, Fahey M, Mahon C, O/'Donoghue AJ, Todorovic A, Morris JH, Maltby DA, Alber T, Cagney G, Bushman FD, Young JA, Chanda SK, Sundquist WI, Kortemme T, Hernandez RD, Craik CS, Burlingame A, Sali A, Frankel AD, Krogan NJ. Global landscape of HIV-human protein complexes. ***Nature***. 2012;**481(**7381):365-70.
30. Au CE, Bell AW, Gilchrist A, Hiding J, Nilsson T, Bergeron JJ. Organellar proteomics to create the cell map. ***Curr Opin Cell Biol***. 2007;**19(**4):376-85. PubMed PMID: 17689063.
31. Yates JR, Gilchrist A, Howell KE, Bergeron JJM. Proteomics of organelles and large cellular structures. ***Nature Reviews Molecular Cell Biology***. 2005;**6(**9):702-14. PubMed PMID: WOS:000231601800012.
32. Taylor SW, Fahy E, Ghosh SS. Global organellar proteomics. ***Trends Biotechnol***. 2003;**21(**2):82-8. PubMed PMID: 12573857.
33. Brennand K, Savas JN, Kim Y, Tran N, Simone A, Hashimoto-Torii K, Beaumont KG, Kim HJ, Topol A, Ladran I, Abdelrahim M, Matikainen-Ankney B, Chao SH, Mrksich M, Rakic P, Fang G, Zhang B, Yates JR, 3rd, Gage FH. Phenotypic differences in hiPSC NPCs derived from patients with schizophrenia. ***Mol Psychiatry***. 2015;**20(**3):361-8. PubMed PMID: 24686136; PubMed Central PMCID: PMC4182344.
34. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, Moest H, Omasits U, Gundry RL, Yoon C, Schiess R, Schmidt A, Mirkowska P, Hartlova A, Van Eyk JE, Bourquin JP, Aebersold R, Boheler KR, Zandstra P, Wollscheid B. A mass spectrometric-derived cell surface protein atlas. ***PLoS One***. 2015;**10(**3):e0121314. PubMed PMID: 25894527; PubMed Central PMCID: PMC4404347.
35. McClatchy DB, Liao LJ, Lee JH, Park SK, Yates JR. Dynamics of Subcellular Proteomes During Brain Development. ***Journal of Proteome Research***. 2012;**11(**4):2467-79. PubMed PMID: ISI:000302388100036.
36. McClatchy DB, Liao LJ, Park SK, Venable JD, Yates JR. Quantification of the synaptosomal proteome of the rat cerebellum during post-natal development. ***Genome Research***. 2007;**17(**9):1378-88. PubMed PMID: ISI:000249236900014.
37. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, Chambers MC, Zimmerman LJ, Shaddox KF, Kim S, Davies SR, Wang S, Wang P, Kinsinger CR, Rivers RC, Rodriguez H, Townsend RR, Ellis MJ, Carr SA, Tabb DL, Coffey RJ, Slebos RJ, Liebler DC. Proteogenomic characterization of human colon and rectal cancer. ***Nature***. 2014;**513(**7518):382-7. PubMed PMID: 25043054; PubMed Central PMCID: PMC4249766.
38. Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villen J, Haas W, Sowa ME, Gygi SP. A tissue-specific atlas of mouse protein phosphorylation and expression. ***Cell***. 2010;**143(**7):1174-89. PubMed PMID: 21183079; PubMed Central PMCID: 3035969.
39. McClatchy DB, Savas JN, Martinez-Bartolome S, Park SK, Maher P, Powell SB, Yates JR, 3rd. Global quantitative analysis of phosphorylation underlying phencyclidine signaling and sensorimotor gating in the prefrontal cortex. ***Mol Psychiatry***. 2015. PubMed PMID: 25869802; PubMed Central PMCID: PMC4605830.
40. Jensen ON. Interpreting the protein language using proteomics. ***Nat Rev Mol Cell Biol***. 2006;**7(**6):391-403. PubMed PMID: 16723975.

41. Kiselar JG, Chance MR. Future directions of structural mass spectrometry using hydroxyl radical footprinting. ***J Mass Spectrom***. 2010;**45(**12):1373-82. PubMed PMID: 20812376; PubMed Central PMCID: PMC3012749.
42. Manjasetty BA, Shi W, Zhan C, Fiser A, Chance MR. A high-throughput approach to protein structure analysis. ***Genet Eng (N Y)***. 2007;**28**:105-28. PubMed PMID: 17153936.
43. Yang X, Wang M, Fitzgerald MC. Analysis of protein folding and function using backbone modified proteins. ***Bioorg Chem***. 2004;**32(**5):438-49. PubMed PMID: 15381405.
44. Iacob RE, Krystek SR, Huang RY, Wei H, Tao L, Lin Z, Morin PE, Doyle ML, Tymiak AA, Engen JR, Chen G. Hydrogen/deuterium exchange mass spectrometry applied to IL-23 interaction characteristics: potential impact for therapeutics. ***Expert Rev Proteomics***. 2015;**12(**2):159-69. PubMed PMID: 25711416; PubMed Central PMCID: PMC4409866.
45. Wales TE, Engen JR. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. ***Mass Spectrom Rev***. 2006;**25(**1):158-70. PubMed PMID: 16208684.
46. Zhang H, Cui W, Gross ML. Mass spectrometry for the biophysical characterization of therapeutic monoclonal antibodies. ***FEBS Lett***. 2014;**588(**2):308-17. PubMed PMID: 24291257; PubMed Central PMCID: PMC3917544.
47. Sadowski PG, Dunkley TP, Shadforth IP, Dupree P, Bessant C, Griffin JL, Lilley KS. Quantitative proteomic approach to study subcellular localization of membrane proteins. ***Nat Protoc***. 2006;**1(**4):1778-89. PubMed PMID: 17487160.
48. Dunkley TP, Watson R, Griffin JL, Dupree P, Lilley KS. Localization of organelle proteins by isotope tagging (LOPIT). ***Mol Cell Proteomics***. 2004;**3(**11):1128-34. PubMed PMID: 15295017.
49. Andersen JS, Wilkinson CJ, Mayor T, Mortensen P, Nigg EA, Mann M. Proteomic characterization of the human centrosome by protein correlation profiling. ***Nature***. 2003;**426(**6966):570-4. PubMed PMID: 14654843.
50. Yates JR, 3rd. Pivotal Role of Computers and Software in Mass Spectrometry - SEQUEST and 20 Years of Tandem MS Database Searching. ***J Am Soc Mass Spectrom***. 2015;**26(**11):1804-13. PubMed PMID: 26286455; PubMed Central PMCID: PMC4625908.

APPENDICES

NSF WORKSHOP: MASS SPECTROMETRY DATA TO KNOWLEDGE

NSF Headquarters Building – Meeting Room 375
4201 Wilson Blvd
Arlington, VA 22230

MEETING AGENDA

Overview of Format The meeting will consist of four topic areas: Data Generation, Data Analysis, Data to Knowledge, and Data Users-Biologists. There will be 2 talks in each topic area. Each talk will be 25 minutes plus 5 minutes for discussion. Each session will be followed by a panel discussion to address specific questions and flesh out areas for future efforts and discussion in the breakout sessions. The last day will start with a summary of the discussions from the previous day to begin the process of drafting a report and recommendations.

WORKSHOP PROGRAM

** Dress will be business casual*
*** Please check-in at the Visitor & Reception Center on the first floor of the NSF Headquarters Building to receive a visitor pass before going on to the meeting room.*

DAY 1: Monday, May 11, 2015

7:45 – 8:00 AM Continental breakfast

8:00-8:10 Dr. Berkowitz NSF – Welcome

1. Mass Spectrometry Data Generation

Questions you address using mass spectrometry and how?
Data analysis workflow used in your studies.
Bottlenecks in data analysis – where people spend their time.

8:10 – 8:40 AM Bottom Up Proteomic Data Generation – Mike Washburn
8:40 – 9:10 AM Top Down Proteomic Data Generation – Ljiljana Paša-Tolić

9:10 – 9:50 AM **Panel Discussion:** Mike W, Ljiljana, Jenny Brodbelt, Michael Wright, MC-Mike MacCoss

- Emerging Data Acquisition Strategies
- Data Analysis Challenges
- Dream experiments- what would you like to do, but can't?

9:50 – 10:00 AM Break

2. Mass Spectrometry Data Analysis - Algorithms and Computation

How flexible are current algorithmic methodologies to handle changing instrumental landscapes – data-volume/size, real-time/on-instrument/in-the-cloud, different analysis modes (fragmentation techniques, resolution improvements, data-independent acquisition, top-down, database search vs. sequence tagging vs. de novo sequencing etc.).
Can current algorithms adapt to a changing computation landscape, e.g. distributed/centralized data, virtualized/real computing, in-memory/out-of-core models, semantic/binary stores, local/distant data, software-as-a-service, and infrastructure-as-a-service.

10:00 – 10:30 AM Bottom Up Data Analysis Workflows – Nathan Edwards
10:30 – 11:00 AM Top Down Data Analysis Workflows – Neil Kelleher
11:00 – 11:45 AM **Panel Discussion**: Nathan, Neil, Marshall, Jeff Agar, Jeff Bilmes, MC-Mark Gerstein

- Changing Computation Landscape
- Emerging computational approaches such as machine learning algorithms
- How to reduce false negatives? (unassigned good spectra)
- How to reduce false discoveries? (false spectrum assignments, false proteins, false knowledge)

11:45 – 12:00 PM Sum Up Morning Session – Yates

12:00 – 1:00 PM Lunch

3. Data to Knowledge

Current methods to translate data to knowledge will be discussed along with limitations and challenges associated with application of these methods to mass spectrometry data sets. Addressing the gap between protein annotation and mass-spectrometry based evidence, e.g. phosphorylation sites, isoforms, occupancy, etc. Knowledge inference from proteomics data repositories: abundance, observability, evidence for true positive protein sequence.

1:00 – 1:30 PM Network Methods to Analyze Large Data Sets – Phil Jaeger
1:30 – 2:00 PM Crowd Sourcing Methods to Annotate Biological Processes – Andra Waagmeester
2:00 – 2:45 PM **Panel Discussion**: Phil Jaeger, Andra Waagmeester, Mark Gerstein, Akhilesh Pandey, MC-Cathy Wu

- Challenges to Integrating MS/Proteomics other Big Data
- Challenges to creating knowledge from MS Big Data
- How to use prior knowledge? (e.g., ProSight's data warehouses),

2:45 – 3:00 PM Break

4. Data Users- Biological Questions to be Answered using MS Data

The goal of this session is to discuss how mass spectrometry data can help solve biological problems and discover mechanisms of disease or biology.
Where does mass spectrometry data fit in biological studies?
What is needed to make it relevant? More IDs? Better Quant? More PTMS?

3:00 – 3:30 PM Biological Case Study I – Ileana Cristea
3:30 – 4:00 PM Biological Case Study II – Steve Briggs
4:00 – 4:45 PM **Panel Discussion:** Ileana, Steve, Peipei, Jenny van Eyk, Cathy Costello, MC-Nuno

- Challenges to biological discovery using MS
- Questions to address using Mass Spectrometry that are not possible now

4:45 – 5:00 PM Sum Up Afternoon Session, charge for Day 2 and break for the night – Yates

6:30 – 7:30 PM Workshop Mixer – Masters Ballroom, Hilton-Arlington

7:30 – 9:30 PM Workshop Dinner – Masters Ballroom, Hilton-Arlington

DAY 2: Tuesday, May 12, 2015

7:45 – 8:00 AM Continental Breakfast

8:00 – 9:45 AM Break Out Groups: 4 groups to formulate important points from the previous day into concrete ideas:
Topic 1 **Mass Spectrometry Data Generation** – Ljiljana Paša-Tolić/Trixi Ueberheide
Topic 2 **MS Data Analysis/Algorithms & Computation –** Nathan Edwards/Nuno Bandeira
Topic 3 **Data to Knowledge** - Marshall Bern/Akhilesh Pandey
Topic 4 **Data Users – Biological Questions to be Answered using MS Data** Peipei Ping/ Jenny Van Eyk

9:45 – 10:00 AM Break

10:00 – 11:30 AM Break Out Groups continue

11:30 – 12:00 PM Summation - John Yates

12:00 PM Adjourn

CONFIRMED PARTICIPANTS

1. **MS Data Generators (10)**

John Yates
Ljiljana Paša-Tolić
Jenny Brodbelt
Benjamin Garcia
Mike Washburn
Neil Kelleher
Trixi Ueberheide
Jeffrey Agar
Michael MacCoss
Catherine Costello

2. **MS Data Analyzers (8)**

Olga Vitek
David Tabb
Jeff Bilmes
Nuno Bandeira
Robert Chalkley
Nathan Edwards
Marshall Bern
Xiuxia Du

3. **Data to Knowledge (Annotators) (6)**

Phil Jaeger
Andra Waagmeester

Mark Gerstein
Cathy Wu
Akhilesh Pandey
Maggie Lam (Ping group)

4. Data Users – Biologists (6)

Jennifery van Eyk
Peipei Ping
Ileana Cristea
Michael Wright
Steve Briggs
Jackie Vogel

5. Note Takers (5)

Lin He (Yates group)
Robin Park (Yates group)
Mathieu Lavallee-Adam (Yates group)
Salvador Martinez de Bartolome Izquierdo (Yates group)
Jordan Kruger (Edwards group)

6. **Observers** (1)

Austin Yang, Ph.D., NIH
Lin He, Ph.D., NSF
Kelsey Cook, Ph.D., NSF
Douglas Sheeley, Ph.D., NIH

WORKSHOP ORGANIZERS

John R. Yates, III, Ph.D.
Ernest W. Hahn Professor
Chemical Physiology and
Molecular and Cellular Neurobiology
10550 North Torrey Pines Road, SR11
The Scripps Research Institute
LaJolla, CA 92037
Email: jyates@scripps.edu

Nathan Edwards, Ph.D.
Associate Professor
Biochemistry and Molecular and Cellular Biology
Georgetown University
37th and O Streets, N.W.,
Washington D.C. 20057
Email: nje5@georgetown.edu

Marshall W. Bern, Ph.D.
Vice President and Research Director
Protein Metrics Inc.
1622 San Carlos Ave. Suite C
San Carlos, CA 94070
Email: bern@proteinmetrics.com

PARTICIPANTS

Jeffrey Agar, Ph.D.
Associate Professor
College of Science
Department of Pharmaceutical Sciences
Northeastern University
Email: j.agar@neu.edu

Nuno Bandeira, Ph.D.
Associate Professor
Dept. Computer Science and Engineering
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of California San Diego
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404, USA
Email: bandeira@ucsd.edu

Jeffrey A. Bilmes, Ph.D.
Department of Electrical Engineering
University of Washington, Seattle
Box 352500
Seattle, WA 98195-2500
bilmes@ee.washington.edu

Steven Briggs, Ph.D.
Distinguished Professor and Chair
Section of Cell and Developmental Biology
6108 Natural Science Building, MC 0380
9500 Gilman Drive
University of California San Diego
La Jolla, CA 92093-0380
Email: SBriggs@ucsd.edu

Jennifer S. Brodbelt, Ph.D.
William H. Wade Professor
Department of Chemistry and Biochemistry
1 University Station A5300
University of Texas at Austin
Austin, TX 78712-0165
Email: jbrodbelt@cm.utexas.edu

Robert Chalkley, Ph.D.
Associate Professor
Department of Pharmaceutical Chemistry
University of California San Francisco
600 16th Street
San Francisco, CA 94143
Email: chalkley@cgl.ucsf.edu

Catherine E. Costello, Ph.D.

William Fairfield Warren Distinguished Professor
Director, Center for Biomedical Mass Spectrometry
Boston University School of Medicine
Mass Spectrometry Resource
670 Albany Street, Rm 511
Boston, MA 02118-2646
Email: cecmsms@bu.edu

Ileana Cristea, Ph.D.
Department of Molecular Biology
Lewis Thomas Laboratory, 210
Washington Road
Princeton, NJ 08544
Email: icristea@princeton.edu

Xiuxia Du, Ph.D.
Associate Professor,
Department of Bioinformatics and Genomics
University of North Carolina at Charlotte
9201 University City Blvd.,
Charlotte, NC 28223
Email: xiuxia.du@uncc.edu

Mark Gerstein, Ph.D.
Williams Professor of Biomedical Informatics
Yale University
MBB, PO Box 208114
New Haven, CT 06520-8114 USA
http://gersteinlab.org
Email mark@gersteinlab.org

Lin He, Ph.D.
Postdoctoral Research Associate
Department of Chemical Physiology
The Scripps Research Institute
10550 North Torrey Pines Road, SR11
La Jolla, CA 92037
Email: linhe43@gmail.com

Philipp Jaeger, Ph.D.
Department of Medicine
University of California San Diego
Email: pjaeger@ucsd.edu

Neil Kelleher, Ph.D.
Walter and Mary Elizabeth Glass Professor in the Life Sciences
Departments of Chemistry, Molecular Biosciences, and the Feinberg School of Medicine
Northwestern University
2145 Sheridan Road, Box #101
Evanston, IL 60208-3113
Phone: 847-467-4362

Email: n-kelleher@northwestern.edu

Jordan Kruger
Graduate Student
Biochemistry and Molecular and Cellular Biology
Georgetown University
37th and O Streets, N.W.,
Washington D.C. 20057
Email: jjk47@georgetown.edu

Maggie P.Y. Lam, Ph.D.
Project Scientist
Department of Physiology
University of California, Los Angeles
675 Charles E. Young Drive
Los Angeles, CA 90095-1760
email: magelpy@ucla.edu

Mathieu Lavallée-Adam, Ph.D.
Postdoctoral Research Associate
Department of Chemical Physiology
The Scripps Research Institute
10550 North Torrey Pines Road, SR11
La Jolla, CA 92037
Email: mlaval@scripps.edu

Michael MacCoss, Ph.D.
Professor of Genome Sciences
Department of Genome Sciences
University of Washington
Seattle, WA
Email: maccoss@u.washington.edu

Salvador Martínez-Bartolomé, Ph.D.
Postdoctoral Research Associate
Department of Chemical Physiology
The Scripps Research Institute
10550 North Torrey Pines Road, SR11
La Jolla, CA 92037
Email: salvador@scripps.edu

Akhilesh Pandey, M.D., Ph.D.
Professor
Institute for Basic Biomedical Sciences
Institute of Genetic Medicine, Departments of Biological Chemistry, Oncology and Pathology
Johns Hopkins University School of Medicine.
Baltimore, MD
Email: pandey@jhmi.edu

Ljiljana Pasa-Tolic, Ph.D.
Lead Scientist for Mass Spectrometry and Group Manager, Mass Spectrometry

Environmental and Molecular Sciences Laboratory
Richland, WA 99354
Email: ljiljana.pasatolic@pnnl.gov

Peipei Ping, Ph.D.
Professor of Physiology, Medicine/Cardiology & Bioinformatics
Director of NIH BD2K Center of Excellence for Biomedical Computing at UCLA
University of California, Los Angeles
675 Charles E. Young Drive
Los Angeles, California 90095
E-mail: pping@mednet.ucla.edu

David L. Tabb, Ph.D.
Department of Biomedical Informatics and Department of Biochemistry
Vanderbilt University School of Medicine
2525 West End Avenue
Suite 1475
Nashville, TN 37203
Email: dtabb1973@gmail.com

Beatrix M Ueberheide, Ph.D.
Assistant Professor,
Department of Biochemistry and Molecular Pharmacology
Director of Proteomics Resource Lab
Rusk, Room 719
400 East 34th Street, New York, NY10016
Email: Beatrix.Ueberheide@nyumc.org

Jennifer Van Eyk, Ph.D.
Erika Glazer Endowed Chair For Women's Heart Health
Director, Basic Science Research, Women's Heart Center
Director, Advanced Clinical Biosystems Research
The Heart Institute
Cedars Sinai Medical Center
Los Angeles, CA
Email: Jennifer.VanEyk@cshs.org

Olga Vitek, Ph.D.
Sy and Laurie Sternberg Interdisciplinary Associate Professor
College of Science
College of Computer and Information Science
360 Huntington Ave.
Boston, Massachusetts 02115
Email: o.vitek@neu.edu

Jackie Vogel, Ph.D.
Director, Integrated Quantitative Biology Initiative (McGill University)
Associate Professor, Department of Biology and the School of Computer Science
McGill University
3649 Promenade Sir William Osler
Montreal QC, H3G 0B1 Canada

email: jackie.vogel@mcgill.ca

Andra Wagmeester, Ph.D.
Data Scientist
Micelio
Antwerp, Belgium
Email: andra@micelio.be

Michael Washburn, Ph.D.
Director of Proteomics Center
Professor, Department of Pathology & Laboratory Medicine
The University of Kansas Medical Center
Stowers Institute for Medical Research
Kansas City, MO
Email: mpw@stowers.org

Michael E. Wright, Ph.D.
Assistant Professor
Molecular Physiology and Biophysics
University of Iowa Carver College of Medicine
University of Iowa
51 Newton Road
Iowa City, IA 52242
Email: michael-e-wright@uiowa.edu

Cathy H. Wu, Ph.D.
Edward G. Jefferson Chair of Bioinformatics & Computational Biology
Director, Center for Bioinformatics & Computational Biology (CBCB)
Director, Protein Information Resource (PIR)
Professor, Computer & Information Sciences
Professor, Biological Sciences
University of Delaware
15 Innovation Way, Suite 205
Newark, DE 19711-5449
Email: wuc@dbi.udel.edu